Computationally Efficient Inference Methods for Biochemical Reaction Networks

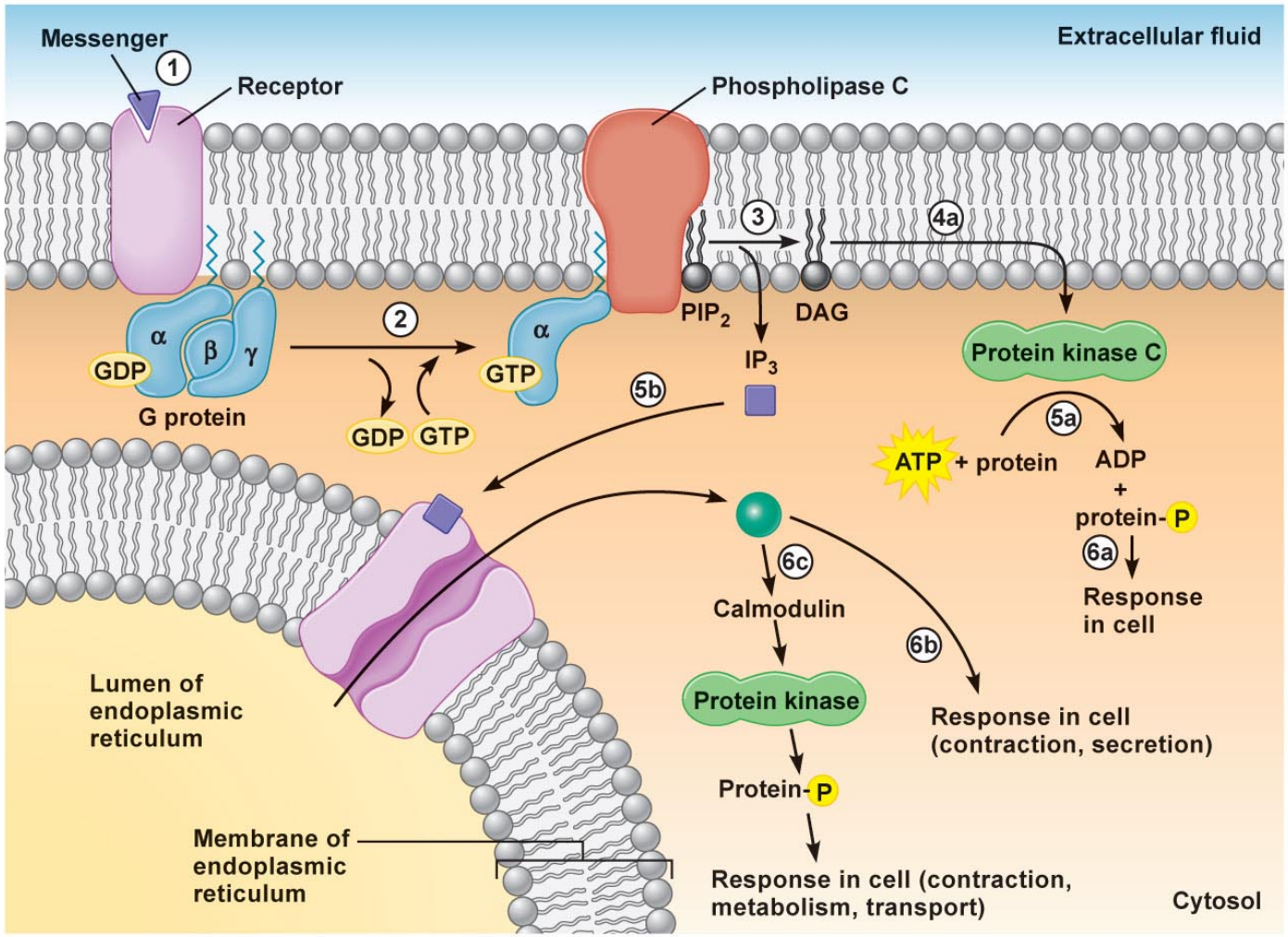
David J. Warne[1], Ruth E. Baker [2], Matthew J. Simpson [1]

[1]School of Mathematical Sciences,
Queensland University of Technology, Australia

[2]Mathematical Institute,
University of Oxford, United Kingdom

11 July 2016

Motivation



1

[1] Stanfield, C. L. *Principles of Human Physiology*, Pearson, 2011

Outline

Background

Biochemical Reaction Network

Let $X_i \in \mathbb{N}$ be the copy numbers of N chemical species, interacting according to the M reactions,

$$\sum_{i=1}^{N} \nu_{i,j}^{-} X_i \xrightarrow{k_j} \sum_{i=1}^{N} \nu_{i,j}^{+} X_i, \quad j = 1, 2, \ldots M.$$

where ν^- and ν^+ are the stoichiometries.

Use stochastic simulation (*Gillespie*[2]) and Monte Carlo methods.

[2] Gillespie, D. T. *The Journal of Physical Chemistry*, 1977

Bayesian Inference

Given $\mathbf{X}_D = [\mathbf{X}(t_1), \mathbf{X}(t_2), \ldots, \mathbf{X}(t_{N_t})]$, what is $\mathbf{k} = [k_1, k_2, \ldots, k_M]$?

Bayes' Theorem

$$p(\mathbf{k} \,|\, \mathbf{X}_D) \propto p(\mathbf{X}_D \,|\, \mathbf{k})\, p(\mathbf{k})$$

The likelihood,

$$p(\mathbf{X}_D \,|\, \mathbf{k}) = \prod_{i=1}^{N_t} p(\mathbf{X}_D(t_i), t_i \,|\, \mathbf{X}_D(t_{i-1}), t_{i-1}, \mathbf{k}),$$

Approximate Bayesian Computation

Basic idea[3]: $p(\mathbf{k} \mid \mathbf{X}_D) \approx p(\mathbf{k} \mid \rho(\mathbf{X}_S, \mathbf{X}_D) \leq \epsilon)$

ABC Rejection Sampling

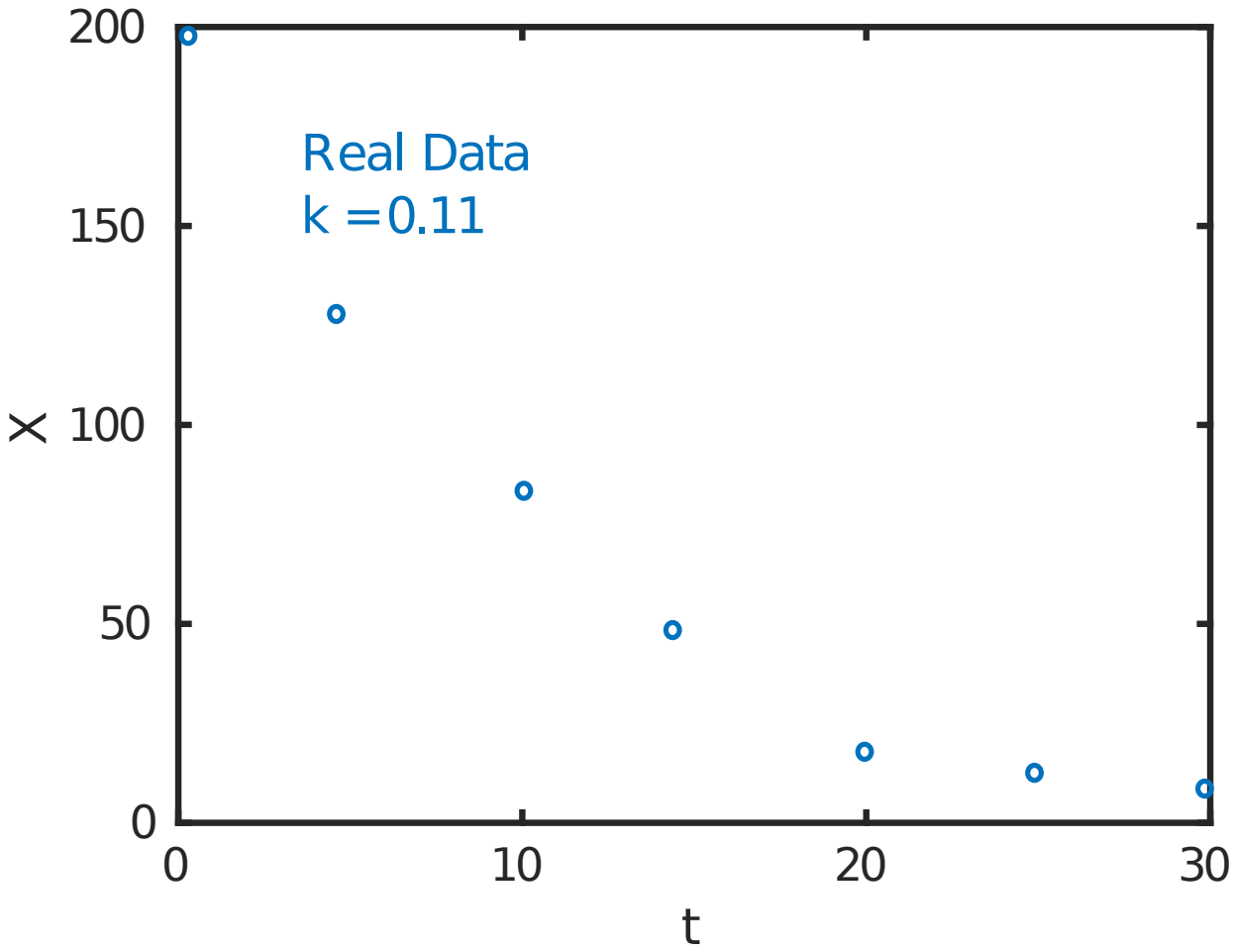
1: **repeat**
2: Sample prior $\mathbf{k}^* \sim p(\mathbf{k})$
3: Generate simulated data $\mathbf{X}_S \sim p(\mathbf{X}|\mathbf{k}^*)$
4: **until** $\rho(\mathbf{X}_S, \mathbf{X}_D) \leq \epsilon$
5: Accept $\mathbf{k}^*$ as a sample from $p(\mathbf{k} \mid \rho(\mathbf{X}_S, \mathbf{X}_D) \leq \epsilon)$

- Compute time depends on the average acceptance rate A_ϵ.
- Unfortunately, $A_\epsilon = O(\epsilon^M)$.

[3] Sunnaker, M., et al. *PLoS Computational Biology*, 2013

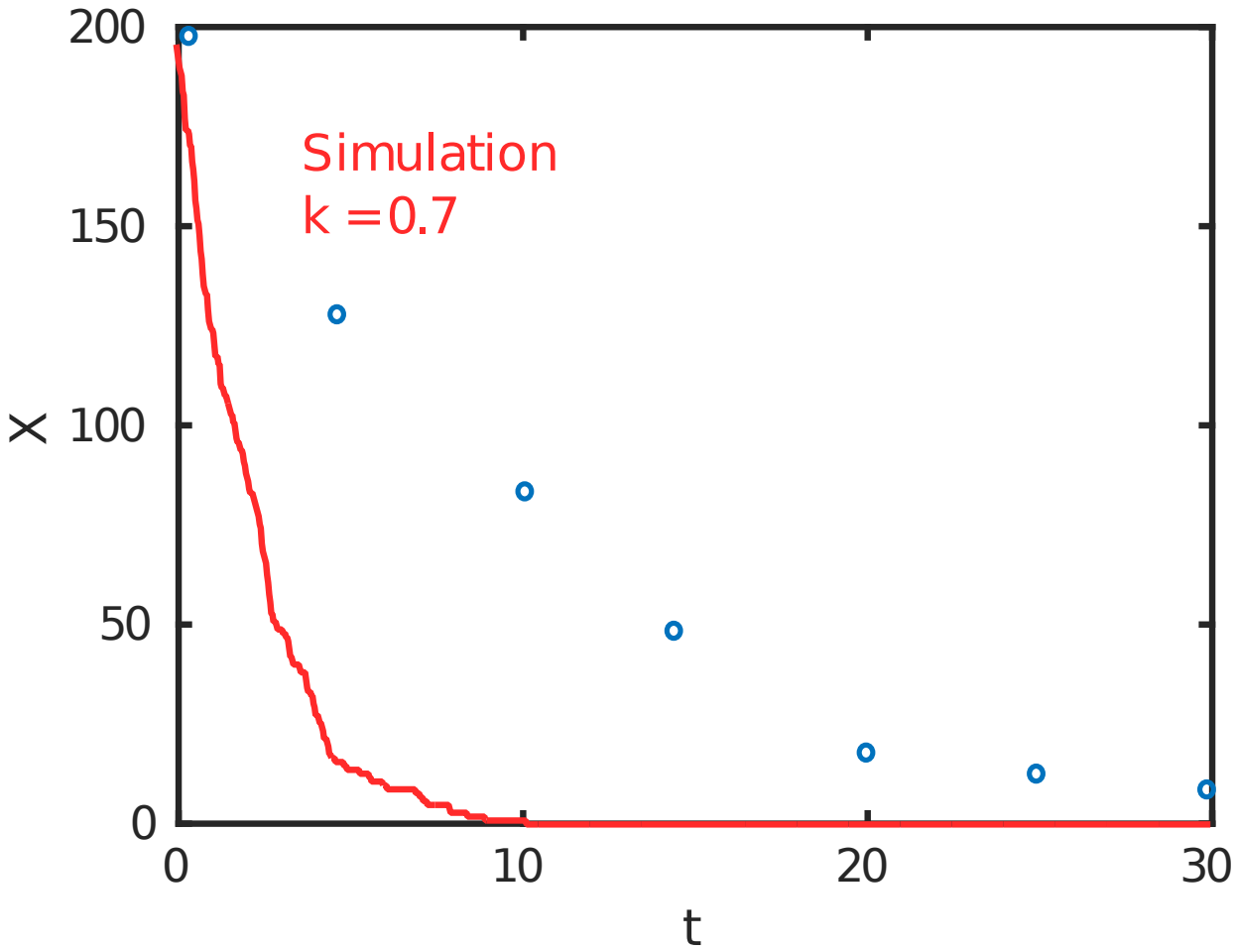
ABC Rejection Sampling Example

$$X \xrightarrow{k} Y$$



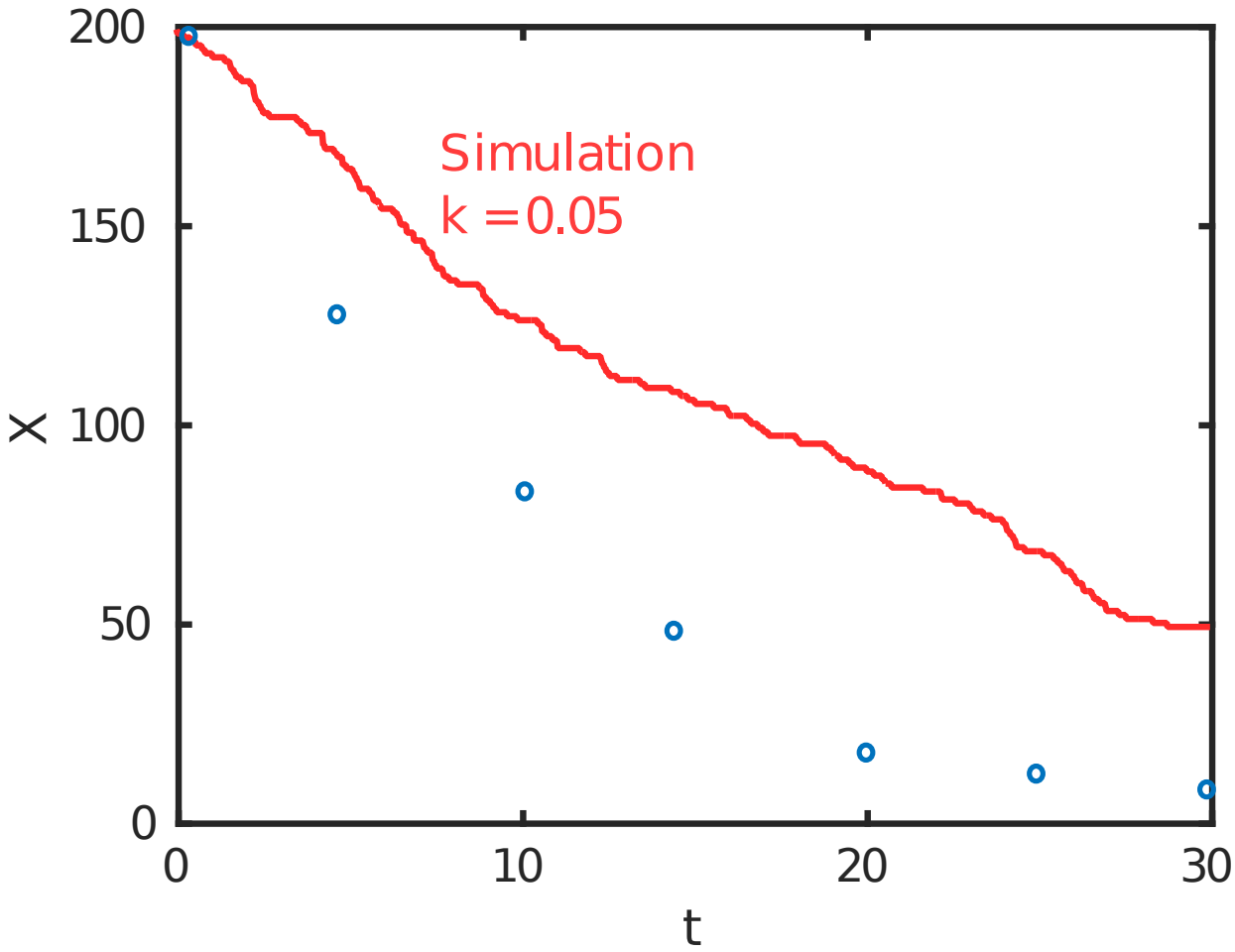
ABC Rejection Sampling Example

$$X \xrightarrow{k} Y$$



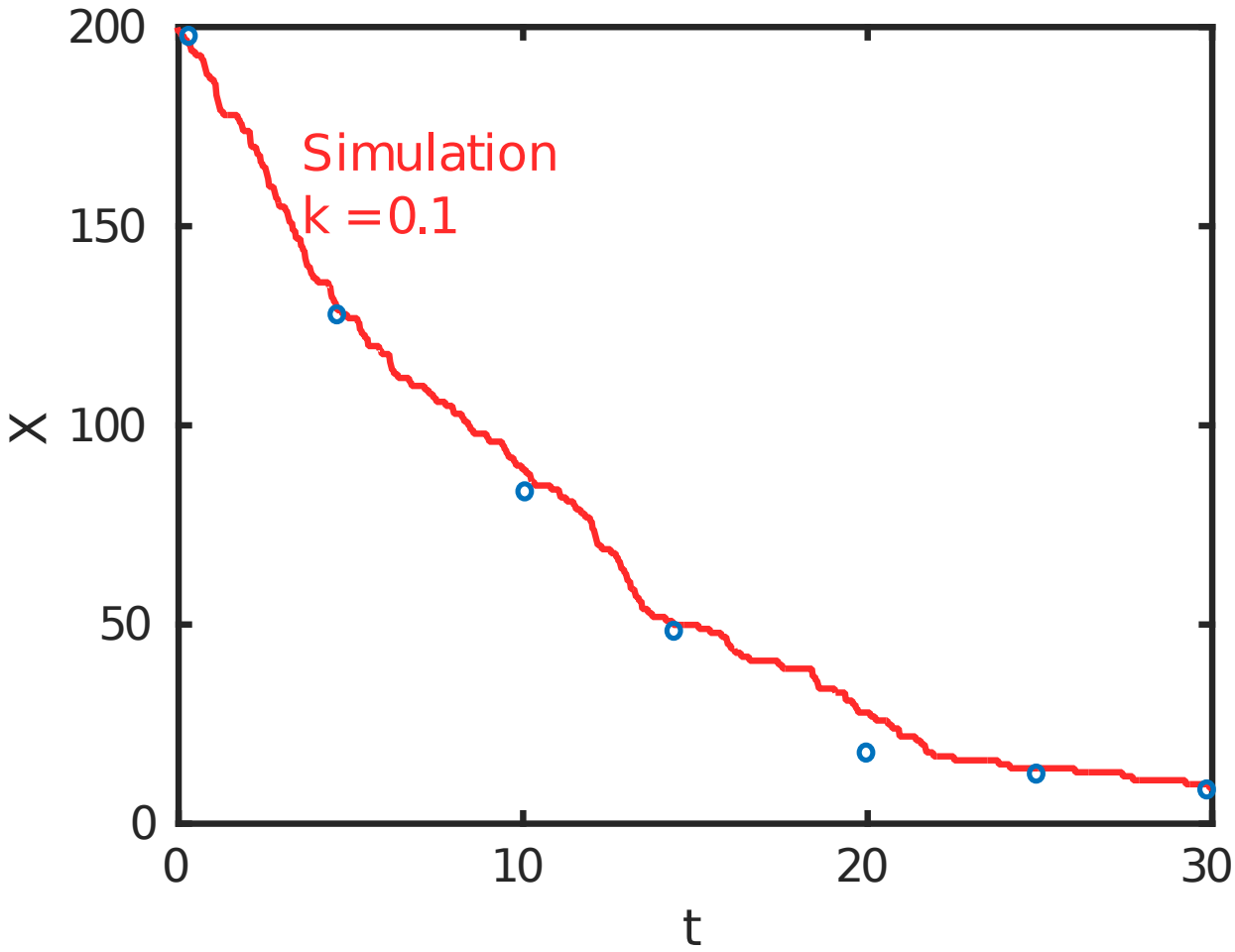
ABC Rejection Sampling Example

$$X \xrightarrow{k} Y$$



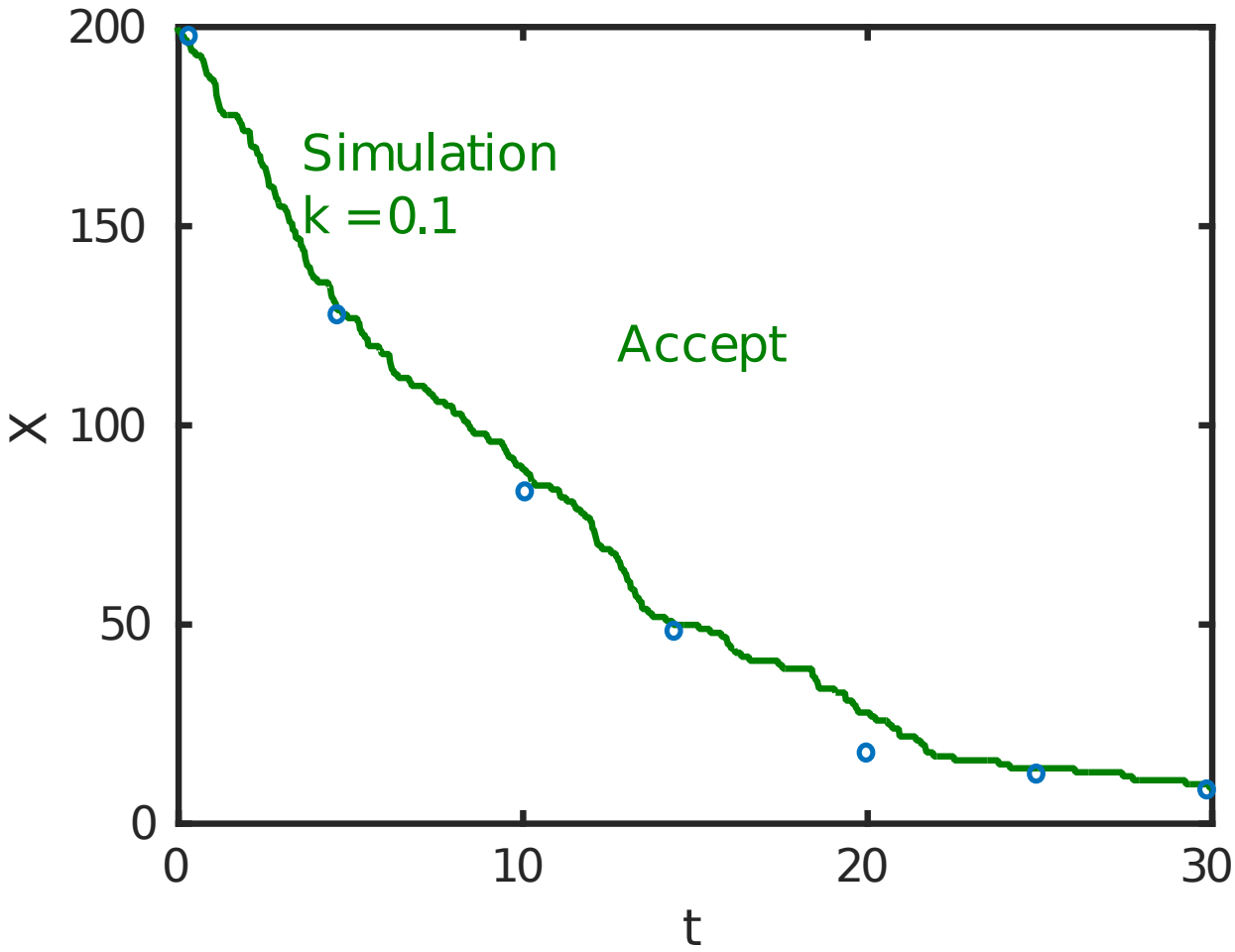
ABC Rejection Sampling Example

$$X \xrightarrow{k} Y$$



ABC Rejection Sampling Example

$$X \xrightarrow{k} Y$$



Multilevel Monte Carlo

Given a random variable, X, with PDF $p(X)$, compute,

$$\mathbb{E}\left[f(X)\right] = \int_{-\infty}^{\infty} f(X)p\left(X\right)\mathrm{d}X.$$

Consider a sequence of approximations [4] $\left\{Y_\ell\right\}_{\ell=0}^{\ell=L}$,

MLMC telescoping sum

$$\mathbb{E}\left[f(X)\right] = \mathbb{E}\left[f(Y_0)\right] + \mathbb{E}\left[f(X) - f(Y_L)\right] + \sum_{\ell=1}^{L} \mathbb{E}\left[f(Y_\ell) - f(Y_{\ell-1})\right],$$

[4] Giles, M. B. *Operations Research*, 2008

Multilevel Approximate Bayesian Computation

Distribution as an Expectation

Let $\boldsymbol{\Theta} = [\theta_1, \theta_2, \ldots, \theta_d]$ be a random vector in $\mathbb{R}^d$. The *cumulative distribution function* (CDF) is given by

$$F(s_1, s_2, \ldots, s_d) = \int_{A(s_1, s_2, \ldots, s_d)} p(\boldsymbol{\Theta}) \; \mathrm{d}\theta_1 \mathrm{d}\theta_2 \ldots \mathrm{d}\theta_d.$$

where $A(s_1, s_2, \ldots, s_d) = (-\infty, s_1] \times (-\infty, s_2] \times \cdots \times (-\infty, s_d]$.

We can reformulate as,

$$F(s_1, s_2, \ldots, s_d) = \mathbb{E}\left[\mathbb{1}_{A(s_1, s_2, \ldots, s_d)}(\boldsymbol{\Theta})\right],$$

Multilevel ABC Estimator

Instead of $p(\mathbf{k} \,|\, \mathbf{X}_D)$, we consider $F(\mathbf{s})$ over $\mathbb{R}^M$.

Multilevel ABC

Given a sequence of ABC acceptance thresholds, $\{\epsilon_\ell\}_{\ell=0}^{\ell=L}$, with $\epsilon_\ell > \epsilon_{\ell+1}$,

$$F(\mathbf{s}) \approx \mathbb{E}\left[\mathbb{1}_{A(\mathbf{s})}(\mathbf{k}_0)\right] + \sum_{\ell=0}^{L} \mathbb{E}\left[\mathbb{1}_{A(\mathbf{s})}(\mathbf{k}_\ell) - \mathbb{1}_{A(\mathbf{s})}(\mathbf{k}_{\ell-1})\right],$$

where $k_\ell \sim p(\mathbf{k} \,|\, \rho(\mathbf{X}_S, \mathbf{X}_D) \leq \epsilon_\ell)$.

Computational Complexity

Key Idea

What the computation time required to estimate the posterior within a desired accuracy?

$$\text{MSE}(\mu) = \mathbb{E}\left[\|F(\mathbf{s}) - \mu(\mathbf{s})\|_\infty^2\right] \leq h^2$$

Standard Monte Carlo,

$$\text{cost}(\mu) = O(h^{-(2+\gamma/\alpha)} \ln h^{-1})$$

Multilevel Monte Carlo,

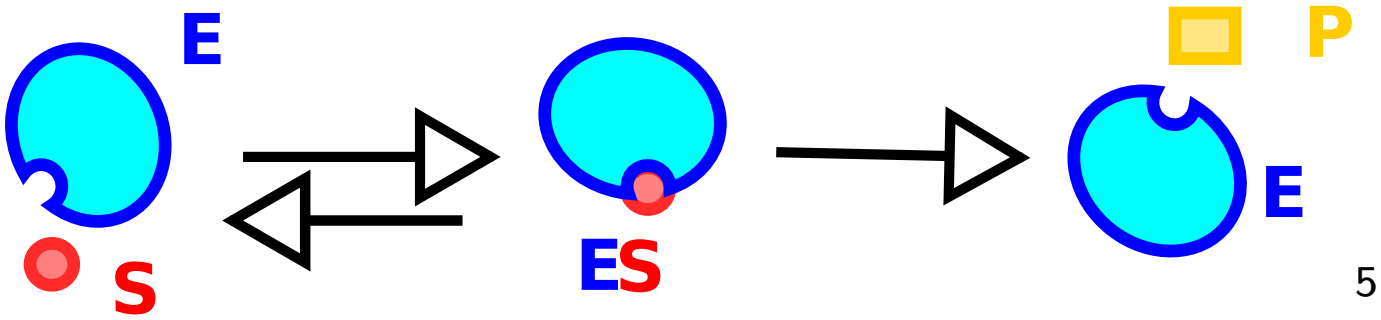
$$\text{cost}(\mu_{ML}) = \begin{cases} O(h^{-2} \ln h^{-1}) & \beta > \gamma \\ O(h^{-2}(\ln h^{-1})^3) & \beta = \gamma \\ O(h^{-(2+(\gamma-\beta)/\alpha)} \ln h^{-1}) & \beta < \gamma \end{cases}$$

Michaelis-Menten Enzyme Kinetics

Michaelis-Menten Model

The catalyzed reaction of substrate, S, to product, P, with enzyme, E, acting as a catalyst.

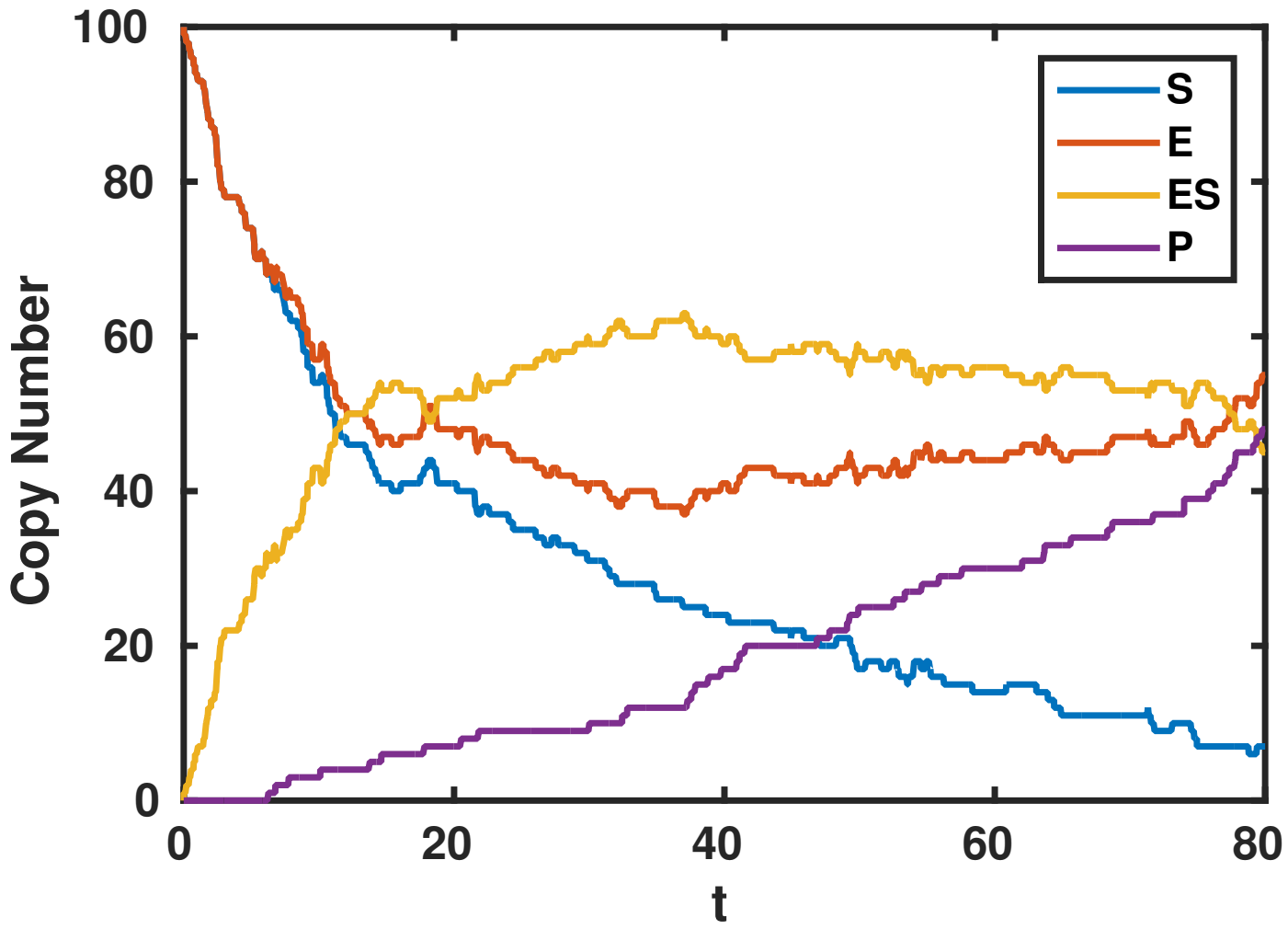
$$S + E \overset{k_1}{\rightarrow} ES, \quad ES \overset{k_2}{\rightarrow} S + E, \quad ES \overset{k_3}{\rightarrow} P + E$$



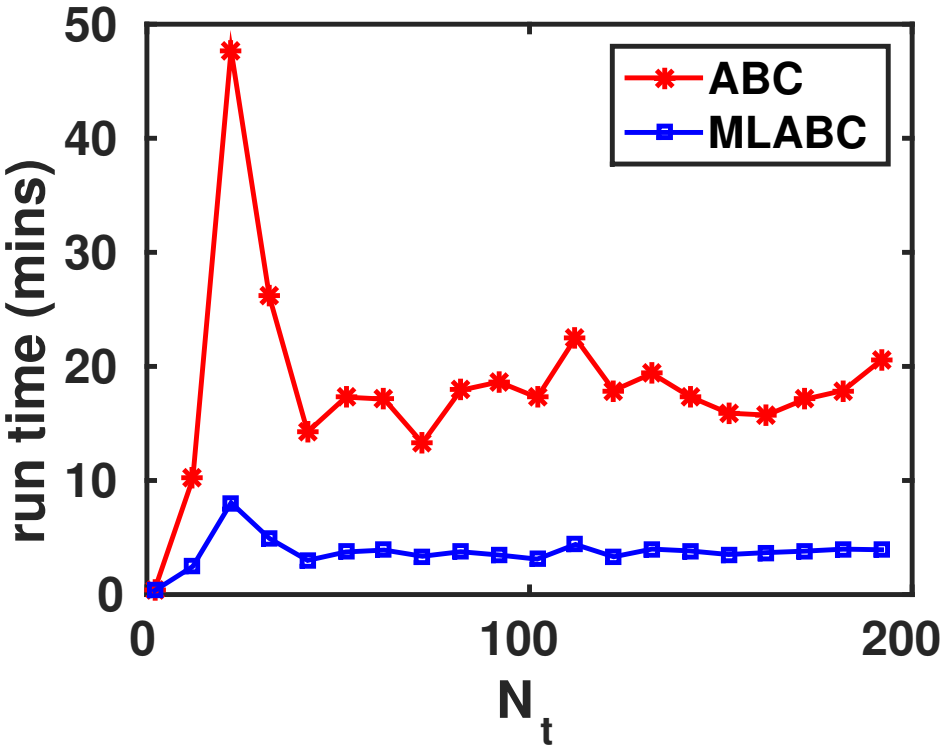
[5]

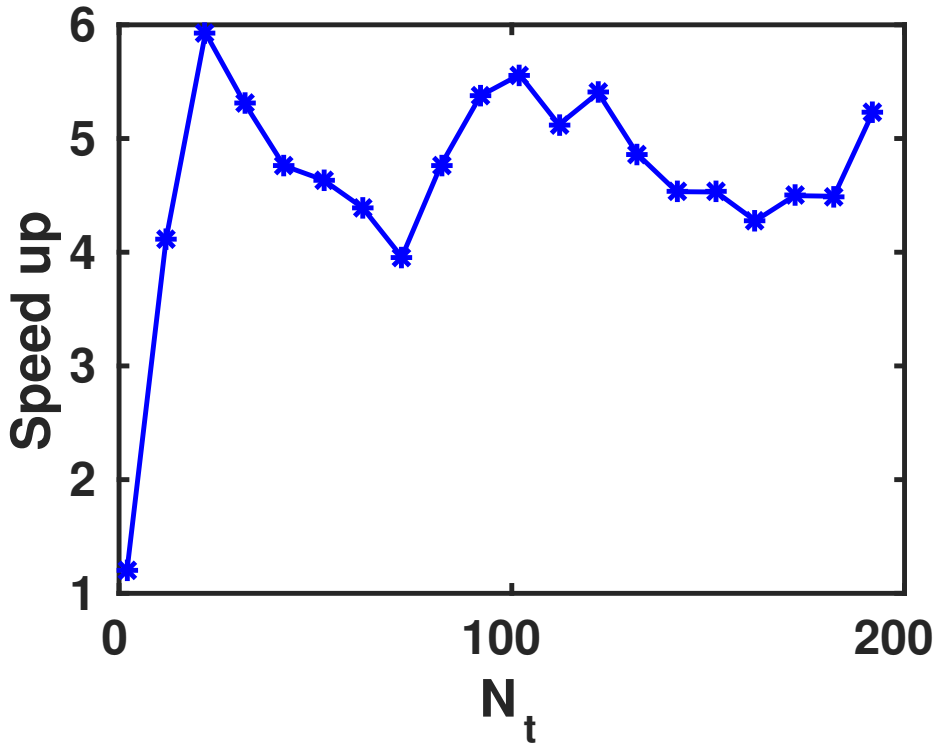
[5] Michaelis, L. and Menten, M. L. *Biochem Z*, 1913

Michaelis-Menten Enzyme Kinetics



Numerical Results: Michaelis-Menten





$h = 0.2$ estimation over $125,000$ points in $\mathbb{R}^3$.

Numerical Results: Michaelis-Menten

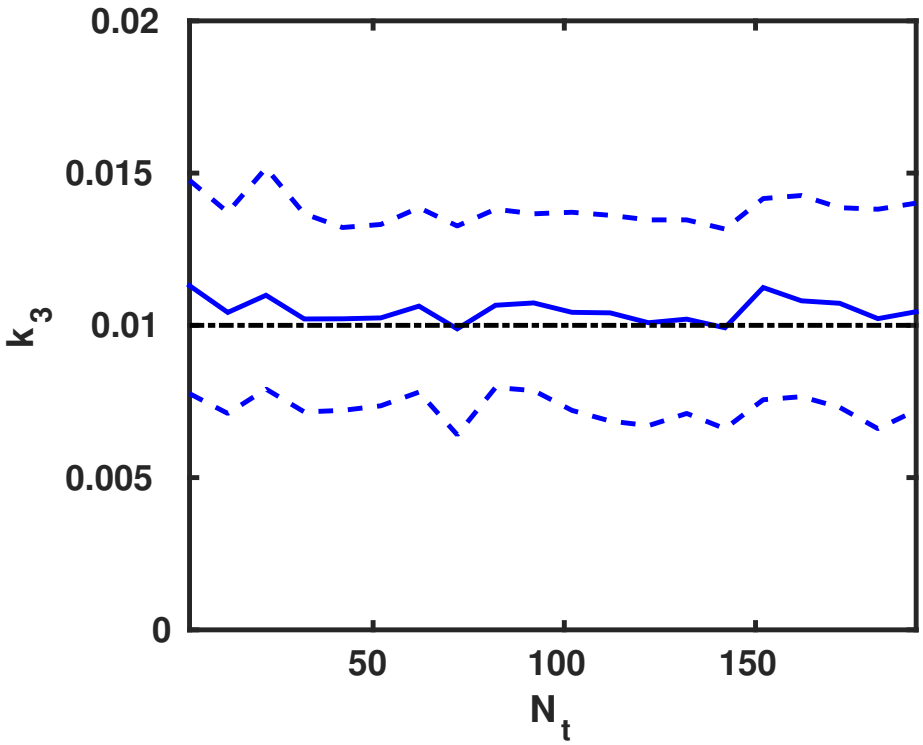


Figure: MLABC

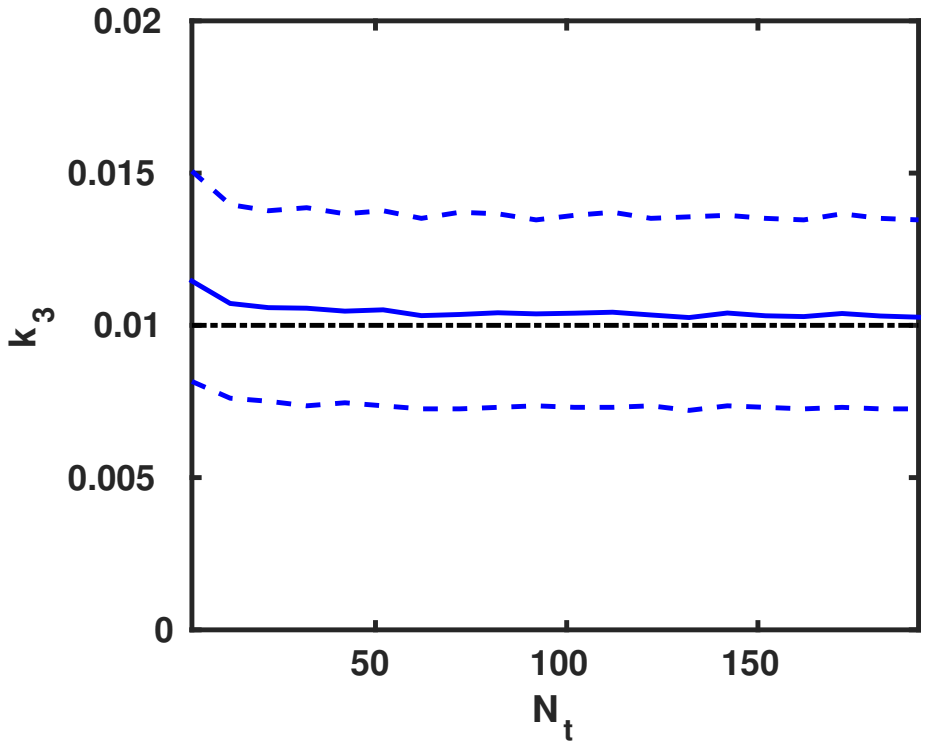


Figure: ABC

The Paper

D. J. Warne, R. E. Baker, M. J. Simpson.
Accelerating computational Bayesian inference for stochastic biochemical reaction network models using multilevel Monte Carlo sampling. Submitting to PLoS Computational Biology.
Soon to be uploaded to bioRxiv.org

Conclusion

- Application of MLMC to ABC inference.
  - General estimator for posterior CDF.
  - Detailed numerical analysis.
  - Promising numerical results.
- Future Work:
  - Further enhancements, e.g., MCMC.
  - Explore the curse-of-dimensionality with MLABC.

Acknowledgments

- PhD Supervisors: Matthew Simpson and Ruth Baker
- Queensland University of Technology (ECMTB2016 funding)
- Society for Mathematical Biology (ECMTB2016 funding)
- Australian government (PhD scholarship)

Thank You!